Compute-Efficient L1 Triage for 12-Hour Kp $\geq$ 5 Geomagnetic-Storm Early Warning

Yeowon Yoon
Troy High School
Fullerton, CA, USA
yeowonyoon0109@gmail.com

***Abstract*—We present a compute-efficient 12-hour geomagnetic storm triage model that uses only upstream solar wind and interplanetary magnetic field (IMF) measurements (L1 or near-Earth) to predict whether storm-level geomagnetic activity (Kp $\geq$ 5, NOAA G1) will occur within the next 12 hours. The system is designed as a front-end gate that triggers costly downstream physics simulations only when risk is elevated. Using NASA OMNI hourly composites spanning 1995-2025 (263,016 timestamps) with strict chronological evaluation under strong class imbalance, we construct supervised datasets from a deployable 10-feature upstream configuration and benchmark multiple time coverage and context variants. Baseline models achieve test ROC-AUC $\approx$ 0.84-0.87 and PR-AUC $\approx$ 0.46-0.52 across variants. On the deployable full-history variant, a tuned random forest reaches ROC-AUC=0.8604 and PR-AUC=0.4839, while a tuned neural network yields higher recall (0.7142 vs. 0.6072 at threshold 0.5) at lower precision, reflecting a recall-first triage policy. Reliability analysis indicates systematic miscalibration at high predicted probabilities, motivating post-hoc calibration and drift aware monitoring prior to operational decision support.**

***Index Terms*—space weather, geomagnetic storms, solar wind, OMNI, Kp index, early warning, triage, random forest, neural network, calibration, successive halving**

I. Introduction

Geomagnetic storms are episodes of efficient solar wind energy transfer into Earth's magnetosphere. Impacts include satellite anomalies, increased drag, navigation error, and geomagnetically induced currents. Coupling is often framed through IMF orientation and reconnection-driven dynamics [1], while empirical coupling functions summarize how upstream solar wind state maps to magnetospheric response [2]. Operationally, major disturbances are frequently associated with CME-driven structures rather than flares alone [3], [4].

Physics-based prediction remains central for Sun-to-Earth forecasting, including background solar wind modeling and CME propagation using WSA/ENLIL-family systems [5]–[7], as well as EUHFORIA and SUSANOO-CME [8], [9]. Running these solvers continuously at high cadence is expensive and, for CME events, sensitive to initialization uncertainty. This paper targets a narrower decision problem: using upstream-only measurements to decide when it is worth paying for downstream simulation.

II. Related Work

Operational pipelines typically combine background solar wind estimates with transient propagation models and post-processing for geomagnetic impact. WSA-style modeling uses near-real-time solar magnetic field updates to improve wind prediction [5], while ENLIL provides a widely used 3D heliospheric MHD framework for solar wind structure and CME propagation [6]. EUHFORIA provides a complementary forecasting environment [8], and recent validation of EUHFORIA's cone and spheromak CME models reports event-set skill and recurring error modes in arrival and Kp-related evaluation [10]. Those results demonstrate a practical result that a functional model is not equally informative in every hour of the solar cycle.

Reduced physics solar wind models such as HUX/HUXt aim to preserve speed while retaining key dynamical behavior [11]. On the machine-learning side, calibration matters whenever scores are treated as probabilities by operators or by automated trigger logic [12]. Under shift, conformal prediction offers conservative uncertainty statements when its assumptions are approximately satisfied [13].

The contribution here is a compute-aware gate, designed to run cheap upstream scoring continuously, then spend compute on downstream solvers during windows where the upstream state looks storm-favorable.

III. Data and Task Definition

A. OMNI Hourly Measurements (1995-2025)

We use NASA SPDF OMNI hourly composites retrieved via OMNIWeb [14], spanning 1995-2025 (263,016 hourly timestamps). OMNI standardizes and time-shifts measurements from multiple spacecraft, which improves usability but introduces multi-decade heterogeneity in measurement provenance and data quality. This motivates chronological evaluation and drift-aware monitoring [15], [16].

B. Inputs and Label Construction

Let X_t denote the upstream measurement vector at hour t. We define a leakage-safe 12-hour warning label using future Kp:

$$y_t = \mathbb{1}\Big(\max_{h\in\{1,\dots,12\}} \mathrm{Kp}_{t+h} \geq 5\Big). \tag{1}$$

Kp is provided by GFZ [17]. In common OMNI exports, Kp appears as Kp$\times$10, so the operational threshold corresponds to $\mathrm{KpIndex} \geq 50$. Kp is never used as an input feature. We retain Dst as contextual metadata (not for labeling), since classic formulations connect ring-current response to upstream driving [18], [19].

C. Leakage Controls

Inputs are restricted to time-t upstream measurements. Labels depend only on future Kp within the fixed 12-hour horizon. Any preprocessing parameters are fit on training data only and applied to validation/test splits.

IV. Feature Set and Dataset Variants

A. Primary 10-Feature Configuration (Deployable)

The deployable configuration (VA) is restricted to ten OMNI upstream solar wind/IMF variables: $|B|$, B_z, B_y (nT); V (km/s), n_p (cm^{-3}), T (K); E_y (mV/m), plasma β, Alfvén Mach number M_A, and dynamic pressure P_{dyn} (nPa). These families follow standard coupling intuition and coupling-function literature [1], [2]. OMNI column names vary by export format; this specification is by physical quantity and unit.

B. Variant Design

We benchmark variants along two axes: (i) time coverage and (ii) feature-family augmentation with context. VA is upstream-only. VB augments VA with solar cycle context (e.g., sunspot number and simple cycle proxies) to test whether coarse cycle state improves discrimination without using magnetospheric indices. VC (VA + Kp as an input) is included only as a diagnostic upper bound; it is excluded from any deployment claims.

C. Missingness Handling

Sentinel values are replaced with NaN. For controlled benchmarking, we remove rows with missing values in the selected feature set, yielding variant-dependent sample sizes. A deployed system would require imputation or sensor fallback; this paper focuses on leakage-safe evaluation with consistent preprocessing.

V. Experimental Protocols

A. Protocol A: Chronological Train/Validation/Test

For each variant, we split chronologically into approximately 66.7% train, 16.7% validation, and 16.7% test with no shuffling.

B. Protocol B: Cross-Cycle Generalization

To isolate cross-cycle generalization, we train on Solar Cycles 23-24 (1996-2019) and evaluate on Cycle 25 (2020-2025) using the VA feature set. This is reported as V10 and is not mixed into Protocol A hyperparameter tuning.

VI. Models and Optimization

We use scikit-learn baselines for tabular prediction [20]. Logistic regression (LR) provides a class-weighted linear baseline, while random forests (RF) provide a non-linear ensemble baseline with feature importance diagnostics. For tuned comparisons on the deployable full-history configuration, we additionally evaluate a small feedforward neural network (NN). NN outputs are treated as scores that may require calibration before probability interpretation [12].

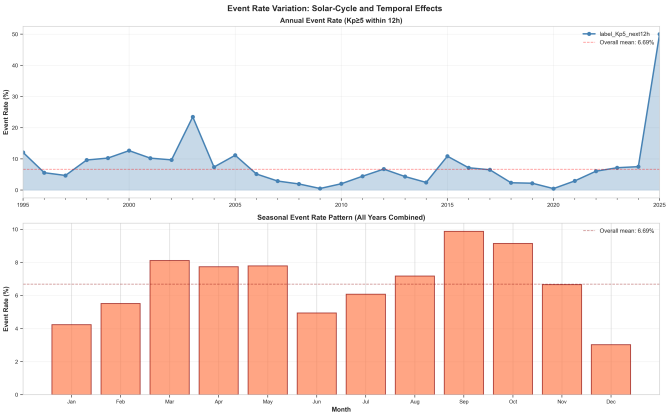

Fig. 1. Event-rate variation over time for the Kp$\geq$ 5 (12-hour) label. To avoid partial year artifacts, exclude incomplete years or annotate coverage (hours observed).

RF tuning uses successive halving via `HalvingRandomSearchCV`, with the number of trees as the resource axis [21]. NN tuning uses an epoch-wise successive halving procedure, promoting candidates with higher validation PR-AUC.

VII. Metrics

Because storm labels are rare and prevalence varies across time windows, we report ROC-AUC for ranking discrimination and PR-AUC for rare-event performance. For a fixed comparison point, we also report recall, precision, and F1 at threshold 0.5. Operational thresholds should be chosen on recent validation windows to satisfy recall targets or trigger-budget constraints.

VIII. Results

A. Event-Rate Context

Event prevalence varies across solar cycle phase, and partial year coverage can inflate annual event-rate estimates if not handled explicitly. Figure 1 visualizes event-rate variation over time for the Kp$\geq$ 5 (12-hour) label and motivates interpreting PR-AUC together with prevalence and observation coverage.

B. Benchmarks Across Variants (Protocol A) and Cross-Cycle Test (Protocol B)

Table I summarizes baseline test ROC-AUC and PR-AUC for LR and RF across ten variants. Across time-coverage windows, ROC-AUC remains comparatively stable, while PR-AUC shifts with prevalence and phase-dependent event density. Figure 2 visualizes this behavior across variants.

C. Storm vs. Non-Storm Separability in Upstream Features

Figure 3 shows representative storm vs. non-storm distributions for IMF variables, supporting that upstream conditions contain separable structure aligned with the label definition. This justifies attempting an upstream-only gate before running propagation models.

Fig. 2. Benchmark performance across dataset variants (ROC-AUC and PR-AUC). PR-AUC varies with prevalence and solar cycle phase.

TABLE I
BENCHMARKING ACROSS 10 VARIANTS. VA/VB ARE DEPLOYABLE (UPSTREAM-ONLY; VB ADDS SOLAR CYCLE CONTEXT). VC INCLUDES KP AS AN INPUT AND IS NON-DEPLOYABLE (DIAGNOSTIC ONLY). V10 IS PROTOCOL B (TRAIN CY23-24, TEST CY25).

Variant	Feat	Train N	LR ROC	LR PR	RF ROC	RF PR	Avg ROC	Comb.
V01_VA_full	10	171254	0.8587	0.4717	0.8663	0.4659	0.86250	0.70502
V02_VA_2000	10	142703	0.8432	0.4781	0.8515	0.4764	0.84735	0.69931
V03_VA_2010	10	86674	0.8553	0.5202	0.8559	0.4944	0.85560	0.71628
V04_VB_full	14	171254	0.8645	0.4750	0.8725	0.4745	0.86850	0.71100
V05_VB_2000	14	142703	0.8492	0.4806	0.8585	0.4908	0.85385	0.70659
V06_VB_2010	14	86674	0.8623	0.5212	0.8679	0.4914	0.86510	0.72158
V07_VC_full†	11	171254	0.8587	0.4717	0.8663	0.4659	0.86250	0.70502
V08_VC_2000†	11	142703	0.8432	0.4781	0.8515	0.4764	0.84735	0.69931
V09_VC_2010†	11	86674	0.8553	0.5202	0.8559	0.4944	0.85560	0.71628
V10_Cycle(VA)	10	210384	0.8583	0.4712	0.8640	0.4642	0.86115	0.70377

†VC uses Kp as an input feature and is shown only as a diagnostic upper bound; it is excluded from deployment claims.

D. Tuned Models on the Deployable Full-History Variant

We select the deployable full-history configuration (V01) as the primary operating point for tuning. RF and NN are tuned on the validation split under Protocol A and evaluated once on the held-out test split. Table II shows that RF improves PR-AUC and precision at the standardized threshold, while NN increases recall. In a gate, the RF behavior reduces wasted triggers, while the NN behavior reduces missed storm windows.

E. Interpretability and Physical Coherence

Figure 4 shows feature importance and dropout sensitivity. Magnetic-field strength and southward IMF components, along with compression proxies such as dynamic pressure, tend to rank highly. This matches the coupling picture where southward B_z and compression help set storm-favorable conditions [1]. Performance degrades gradually under ablations, suggesting the model is not dominated by a single brittle input.

F. Calibration and Reliability

Figure 5 shows reliability curves indicating miscalibration in the high-score region. If scores are read as literal probabilities, overconfidence can cause trigger policies to behave too aggressively. In deployment, outputs should be treated as scores unless calibrated on recent windows using post-hoc methods [12]. Under distribution shift, conformal methods can add a conservative uncertainty layer when assumptions are approximately met [13].

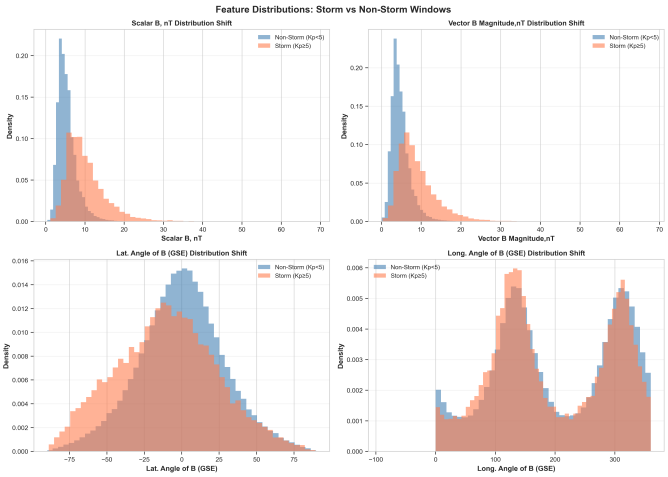

Fig. 3. Storm vs. non-storm feature distributions for representative IMF variables, supporting physical separability of upstream conditions.

TABLE II
TUNED TEST-SET PERFORMANCE ON V01 (VA_FULL, 1995-2025). THRESHOLDED METRICS AT 0.5.

Model	PR-AUC	ROC-AUC	Recall	Precision	F1
RF (tuned)	0.4839	0.8604	0.6072	0.3460	0.4408
NN (tuned)	0.4626	0.8521	0.7142	0.1979	0.3100

IX. OPERATIONAL CONTEXT AND DEPLOYMENT INTERPRETATION

This model serves as a trigger and does not forecast CME propagation. It answers a narrower evaluation of the cost of running a heavier model in the next cycle.

Figure 2 suggests the rank-order signal is stable across time windows, while Figure 5 shows that probability calibration needs separate attention. Figure 3 supports that upstream separability exists, and Figure 4 shows that the strongest predictors align with well-known storm drivers.

EUHFORIA validation results make the compute-allocation argument concrete. In event sets, skill and error vary by configuration and by case [10]. Running a full CME model every hour is not realistic for many settings, and it is not necessary during quiet upstream regimes. A gate makes the expensive model run often when upstream risk is elevated and less often when conditions look benign. When escalation is triggered, reduced-physics fast models (e.g., HUXt) can be used as cheap bracketing before spending compute on higher-fidelity ensembles [11].

X. DATA SHIFT, EVENT-RATE VARIABILITY, AND DRIFT

Storm prevalence varies across cycle phase and season. Figure 1 shows that the event rate is not stationary, so PR-AUC

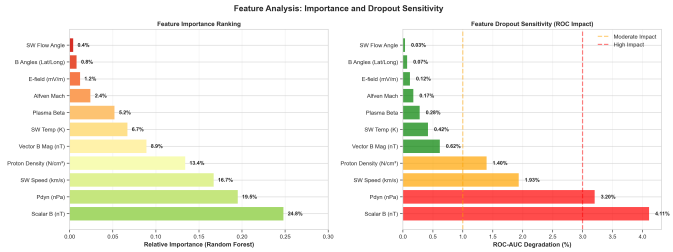

Fig. 4. Feature importance and dropout sensitivity. Dominant predictors commonly include $|B|$, B_z, P_{dyn}, and solar wind speed, consistent with coupling intuition [1].

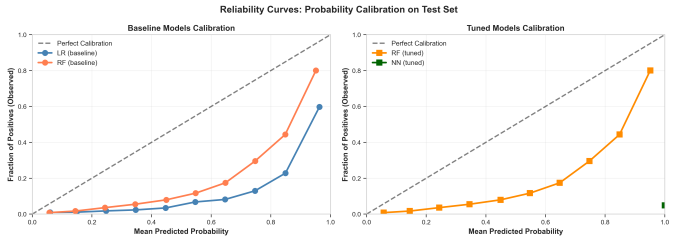

Fig. 5. Reliability curves (calibration). Curves below the diagonal indicate overconfident probabilities, motivating post-hoc calibration prior to operational use [12].

should be read together with prevalence and data coverage. Because OMNI is a composite record and measurement provenance changes across decades, distribution shift is expected [15]. In deployment this motivates rolling validation windows, periodic recalibration, and drift monitoring that is driven by recent data rather than pooled multi-decade averages.

XI. Limitations

The label is defined using Kp within a fixed 12-hour window. This matches a NOAA G1-aligned triage objective, but it does not cover all hazards. Alternative targets (Dst thresholds, sustained southward B_z, AE bursts) correspond to different operational definitions and may change which upstream features dominate. Missingness is handled by row deletion for controlled benchmarking; real deployments require imputation or fallback logic. Calibration remains imperfect in high-score bins (Figure 5), so probability interpretation requires post-hoc calibration on recent data and continuous monitoring [12].

XII. Conclusion

We developed a compute-efficient 12-hour geomagnetic-storm triage model that predicts whether Kp ≥ 5 will occur within the next 12 hours using only upstream solar wind and IMF measurements. With leakage-safe labels and strict chronological evaluation, discrimination is stable across time-coverage and context variants (Figure 2), while PR-AUC moves with event prevalence (Figure 1). On the deployable full-history configuration, a tuned random forest improves PR-AUC and precision, while a tuned neural network increases recall, reflecting two plausible gate settings. Feature separability (Figure 3) and physically aligned predictor rankings (Figure 4) support an upstream-only trigger, while reliability

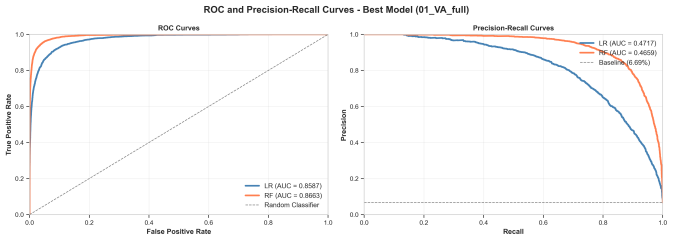

Fig. 6. ROC and PR curves for representative models on the deployable configuration.

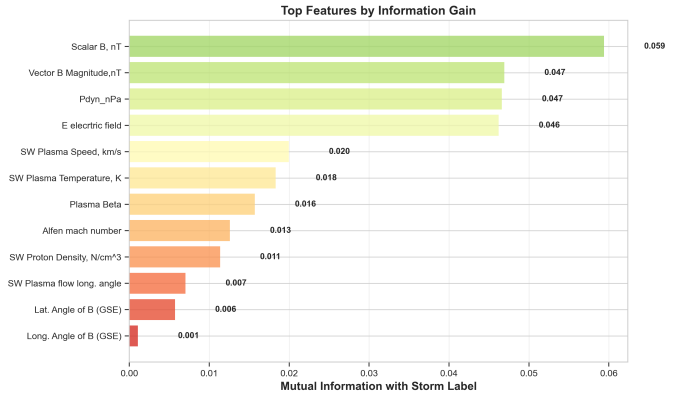

Fig. 7. Mutual-information feature ranking (model-agnostic).

curves (Figure 5) show that calibration must be handled explicitly before operational use.

Acknowledgments

The author thanks mentor on experimental design, evaluation protocol, and scientific writing.

Appendix A
Additional Figures

The appendix provides supporting plots that are useful for interpretation and reproducibility but not required for the main narrative.

References

[1] J. W. Dungey, "Interplanetary magnetic field and the auroral zones," *Physical Review Letters*, vol. 6, no. 2, pp. 47–48, 1961.
[2] P. T. Newell, T. Sotirelis, K. Li, C.-I. Meng, and F. J. Rich, "A nearly universal solar wind–magnetosphere coupling function inferred from ten magnetospheric state variables," *Journal of Geophysical Research: Space Physics*, vol. 113, p. A04218, 2008.
[3] J. T. Gosling, "The solar flare myth," *Journal of Geophysical Research: Space Physics*, vol. 98, no. A11, pp. 18 937–18 949, 1993.
[4] D. F. Webb and T. A. Howard, "Coronal mass ejections: Observations," *Living Reviews in Solar Physics*, vol. 9, no. 3, 2012.
[5] C. N. Arge and V. J. Pizzo, "Improvement in the prediction of solar wind conditions using near-real time solar magnetic field updates," *Journal of Geophysical Research: Space Physics*, vol. 105, no. A5, pp. 10 465–10 479, 2000.
[6] D. Odstrčil, "Modeling 3-D solar wind structure," *Advances in Space Research*, vol. 32, no. 4, pp. 497–506, 2003.
[7] "Noaa swpc: Wsa–enlil solar wind prediction product," https://www.swpc.noaa.gov/, accessed 2026-01-19.

Fig. 8. Train-validation generalization gaps under chronological splits.

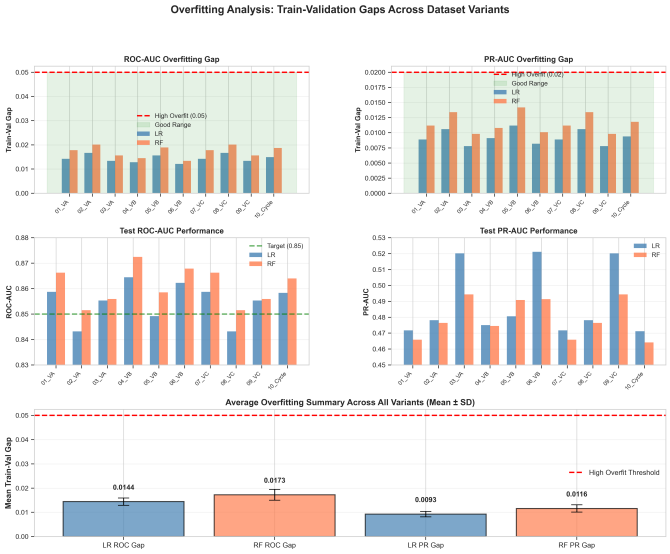

Fig. 9. Summary grid of generalization gaps and test metrics across variants.

[8] J. Pomoell and S. Poedts, "EUHFORIA: European heliospheric forecasting information asset," *Journal of Space Weather and Space Climate*, vol. 8, p. A35, 2018.

[9] D. Shiota and R. Kataoka, "SUSANOO-CME: MHD simulation of interplanetary propagation of multiple CMEs with internal magnetic flux rope," *Space Weather*, vol. 14, pp. 56–75, 2016.

[10] L. Rodríguez, D. Shukhobodskaia, A. Niemela, A. Maharana, E. Samara, C. Verbeke, J. Magdalenić, R. Vansintjan, M. Mierla, C. Scolini, R. Sarkar, E. Kilpua, E. Asvestari, K. Herbst, G. Lapenta, A. D. Chaduteau, J. Pomoell, and S. Poedts, "Validation of EUHFORIA cone and spheromak coronal mass ejection models," *Astronomy & Astrophysics*, 2024.

[11] L. Barnard, M. J. Owens *et al.*, "HUXt—an open-source, reduced-physics solar wind model," *Frontiers in Physics*, vol. 10, 2022.

[12] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[13] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Foundations and Trends in Machine Learning*, vol. 16, no. 4, 2023.

[14] "Nasa spdf omniweb: Omni2 data access," https://omniweb.gsfc.nasa.gov/, accessed 2026-01-19.

[15] J. H. King and N. E. Papitashvili, "Solar wind spatial scales in and comparisons of hourly Wind and ACE plasma and magnetic field data," *Journal of Geophysical Research: Space Physics*, vol. 110, p. A02104, 2005.

[16] R. P. Lepping, M. H. Acuña *et al.*, "The Wind magnetic field investigation," *Space Science Reviews*, vol. 71, pp. 207–229, 1995.

[17] "Gfz potsdam: Geomagnetic Kp index," https://kp.gfz.de/, accessed 2026-01-19.

[18] R. K. Burton, R. L. McPherron, and C. T. Russell, "An empirical relationship between interplanetary conditions and Dst," *Journal of Geophysical Research*, vol. 80, no. 31, pp. 4204–4214, 1975.

[19] "Wdc for geomagnetism, kyoto: Dst index service," https://wdc.kugi.kyoto-u.ac.jp/dstdir/, accessed 2026-01-19.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] "Scikit-learn documentation: Successive halving (HalvingRandomSearchCV)," https://scikit-learn.org/stable/modules/grid_search.html#successive-halving-user-guide, accessed 2026-01-19.